CMO: Quasi-Newton Methods

Eklavya Sharma

Contents

1 Quasi-Newton method template

Newton's method's update rule:

$$x_{k+1} = x_k - \mathrm{H}_f^{-1}(x_k)\,\nabla_f(x_k)$$

This method is not useful, because it requires inverting the hessian, which can be prohibitively computationally expensive for high-dimensional data.

We will therefore try to model the change in the hessian's inverse, and approximate the hessian's inverse instead of calculating it exactly.

Let $g_k = \nabla_f(x_k)$, $\delta_k = x_{k+1} - x_k$ and $\gamma_k = g_{k+1} - g_k$.

$$\nabla_f(x_{k+1}) \approx \nabla_f(x_k) + \mathrm{H}_f(x_k)(x_{k+1} - x_k) \qquad \text{(by differentiating Taylor series)}$$

$$\implies \delta_k \approx \mathrm{H}_f^{-1}(x_k)\gamma_k$$

This inspires us to use an update rule of this form:

$$x_{k+1} = x_k - A_k g_k$$

and apply the following constraint on A_k:

$$\delta_k = A_{k+1}\gamma_k \tag{1}$$

This constraint is called the ‘Quasi-Newton condition’.

Also, we must ensure that A_k is symmetric and positive (semi)definite.

Note that the Quasi-Newton condition is d equations, whereas there are d^2 entries in A_k. We therefore have a lot of slack in terms of how to update A_k.

In all Quasi-Newton methods described next, we choose A_0 as any matrix which is symmetric and positive (semi)definite. Generally, the identity matrix is used. Then we use A_k, δ_k and γ_k to obtain A_{k+1} via an update rule, like ‘rank-1 update’, ‘rank-2 update’ or ‘BFGS’.

2 Rank-1 update

Here we impose a condition of the form $A_{k+1} = A_k + cuu^T$, where $c \in \mathbb{R}$ and $u \in \mathbb{R}^d$ (Note that $\operatorname{rank}(uu^T) = 1$).

It’s easy to see that A_{k+1} is symmetric for all c and positive definite for $c \geq 0$.

To get concrete values of c and u, we’ll plug the rank-1 update condition into the Quasi-Newton condition (1).

$$\delta_k = (A_k + cuu^T)\gamma_k \implies (cu^T\gamma_k)u = \delta_k - A_k\gamma_k$$

Therefore, u is parallel to $\delta_k - A_k\gamma_k$. Let $u = \delta_k - A_k\gamma_k$. Then

$$u = (cu^T\gamma_k)u \implies cu^T\gamma_k = 1 \implies c = \frac{1}{u^T\gamma_k} = \frac{1}{\delta_k^T\gamma_k - \gamma_k^T A_k\gamma_k}$$

With these specific values of u and c, the rank-1 update condition will satisfy all required conditions (symmetry, positive definiteness and Quasi-Newton condition) if $c \geq 0$.

Unfortunately, it has not yet been proved or disproved whether $c \geq 0$.

2.1 Analysis for quadratic function

Let $f(x) = \frac{1}{2}x^TQx - b^Tx$, where Q is symmetric and positive definite. Then $\nabla_f(x) = Qx - b \implies \gamma_k = Q\delta_k$.

Lemma 1.

$$\forall i \in [0, k], A_{k+1}\gamma_i = \delta_i$$

Proof by induction on k.

$$P(l) : \forall i \in [0, l-1], A_l\gamma_i = \delta_i$$

We have to prove $P(l)$ for all $l \geq 1$.

Base case: Since A_1 was constructed to follow the Quasi-Newton condition, $\delta_0 = A_1\gamma_0 \implies P(1)$.

Inductive step: Assume $P(l)$ is true. We’ll prove $P(l+1)$.

Let $i \in [0, l-1]$.

$$\begin{aligned} A_{l+1}\gamma_i &= \left(A_l + \frac{uu^T}{u^T\gamma_l}\right)\gamma_i && \text{(here } u = \delta_l - A_l\gamma_l\text{)} \\ &= \delta_i + \frac{u^T\gamma_i}{u^T\gamma_l}u && (A_l\gamma_i = \delta_i \text{ by induction hypothesis}) \end{aligned}$$

$$\begin{aligned} u^T\gamma_i &= (\delta_l - A_l\gamma_l)^T\gamma_i \\ &= \delta_l^T\gamma_i - \gamma_l^T A_l\gamma_i \\ &= \delta_l^T\gamma_i - \gamma_l^T\delta_i && \text{(by induction hypothesis)} \\ &= \delta_l^T Q\delta_i - \delta_l^T Q\delta_i && (\forall j, \gamma_j = Q\delta_j) \\ &= 0 \end{aligned}$$

Therefore, $A_{l+1}\gamma_i = \delta_i$ for all $i \in [0, l-1]$. Since A_{l+1} was constructed to follow the Quasi-Newton condition, $A_{l+1}\gamma_l = \delta_l$. Therefore, $P(l+1)$ holds true. □

Lemma 2. *If all δ_i were orthonormal, then $A_d = Q^{-1}$.*

Proof. By lemma 1,

$$\forall i \in [0, d-1], \delta_i = A_d\gamma_i = A_dQ\delta_i$$

Therefore, $(1, \delta_i)$ is an eigenpair for A_dQ.

Let P be the matrix whose i^{th} columns is δ_i. P exists because real symmetric matrices are orthogonally diagonalizable and A_dQ is real and symmetric. Then $A_dQ = PIP^T = I \implies A_d = Q^{-1}$. □

Lemma 3. *If all δ_i are linearly independent, then $A_d = Q^{-1}$.*

Proof. Let $\Delta = \{\delta_0, \ldots, \delta_{d-1}\}$. Since $\Delta \subseteq \mathbb{R}^d$, $|\Delta| = d = \dim(\mathbb{R}^d)$ and Δ is linearly independent, Δ is a basis of $\mathbb{R}^d$.

Let $x \in \mathbb{R}^d$. Let $x = \sum_{i=0}^{d-1} c_i\delta_i$. Then

$$A_dQx = \sum_{i=0}^{d-1} A_dQ(c_i\delta_i) = \sum_{i=0}^{d-1} c_i(A_d\gamma_i) = \sum_{i=0}^{d-1} c_i\delta_i = x$$

Therefore, $\forall x \in \mathbb{R}^d, (A_dQ)x = x$, so $A_dQ = I$.

Note that the proof is not specific to rank-1 updates. Its correctness relies only on the Quasi-Newton condition and f being quadratic. □

Since $A_d = Q^{-1}$, the $(d+1)^{\text{th}}$ iteration would be identical to Newton's method. So the rank-1 update method will converge to the minimum in at most $d+1$ iterations.

2.2 Unresolved questions

- A_k is positive definite when $c \geq 0$. Is $c \geq 0$?
- Is $\{\delta_0, \delta_1, \ldots\}$ linearly independent?

3 Rank-2 update

$$A_{k+1} = A_k + cuu^T + bvv^T$$

It's easy to see that A_{k+1} is symmetric iff A_k is symmetric.

By Quasi-Newton condition, we get

$$\delta_k = A_{k+1}\gamma_k \implies (cu^T\gamma_k)u + (bv^T\gamma_k)v = \delta_k - A_k\gamma_k$$

Let $u = \delta_k$ and $v = A_k\gamma_k$. Then

$$c = \frac{1}{u^T\gamma_k} = \frac{1}{\delta_k^T\gamma_k} \qquad b = \frac{-1}{v^T\gamma_k} = \frac{-1}{\gamma_k^T A_k\gamma_k}$$

$$A_{k+1} = A_k + \frac{\delta_k\delta_k^T}{\delta_k^T\gamma_k} - \frac{A_k\gamma_k\gamma_k^T A_k}{\gamma_k^T A_k\gamma_k}$$

3.1 Analysis for quadratic function

Let $f(x) = \frac{1}{2}x^TQx - b^Tx$. Then $\gamma_k = Q\delta_k$.

Lemma 4 (Symmetric square root of a matrix)**.** *If A is a symmetric and positive definite matrix, then $\exists L$ such that $A = L^2$ and L is symmetric, positive semidefinite and invertible.*

Proof. Since A is real and symmetric, it is orthogonally diagonalizable. So there is a matrix P and a diagonal matrix D such that $A = PDP^T$ and $PP^T = P^TP = I$. Since A is positive definite, all diagonal entries of D are positive. Therefore, $\sqrt{D}$ exists. Also, all entries of $\sqrt{D}$ are positive, so $\sqrt{D}^{-1}$ exists. Let $L = P\sqrt{D}P^T$. Then L is symmetric and $L^2 = A$.

$$u^TL^u = u^T(P\sqrt{D}P^T)u = (P^Tu)^T\sqrt{D}(P^Tu) \geq 0$$

Therefore, L is also positive semidefinite. Also,

$$L(P\sqrt{D}^{-1}P^T) = P\sqrt{D}P^TP\sqrt{D}^{-1}P^T = I$$

Therefore, $L^{-1} = P\sqrt{D}^{-1}P^T$. □

Theorem 5. *Let A_k be symmetric and positive definite. Then A_{k+1} is positive definite.*

Proof.

$$c = \frac{1}{\delta_k^T\gamma_k} = \frac{1}{\delta_k^TQ\delta_k} > 0 \tag{2}$$

We'll now prove that $A_{k+1} - cuu^T$ is positive semidefinite. Let $w \in \mathbb{R}^d - \{0\}$.

$$\begin{aligned} &w^T(A_{k+1} - cuu^T)w \\ &= w^T(A_k + bvv^T)w \\ &= w^TA_kw - \frac{(w^TA_k\gamma_k)^2}{\gamma_k^TA_k\gamma_k} \end{aligned}$$

Since A_k is symmetric and positive definite, it has a symmetric and invertible square root L.

$$\begin{aligned}
&w^T(A_{k+1} - cuu^T)w \\
&= w^T L^T L w - \frac{(w^T L^T L \gamma_k)^2}{\gamma_k^T L^T L \gamma_k} \\
&= \|Lw\|^2 - \frac{((Lw)^T(L\gamma_k))^2}{\|L\gamma_k\|^2} \\
&\geq 0 && \text{(by Cauchy-Schwarz inequality)}
\end{aligned}$$

Therefore, $A_{k+1} - cuu^T$ is positive semidefinite. Since cuu^T is also positive semidefinite, A_{k+1} is also positive semidefinite.

The Cauchy-Schwarz inequality is tight iff the vectors are parallel or anti-parallel. Therefore, $A_{k+1} - cuu^T = 0 \iff Lw = \alpha L\gamma_k$ for some $\alpha \in \mathbb{R}$. Since L is invertible, this is equivalent to $w = \alpha\gamma_k$.

Assume A_{k+1} is not positive definite. $\exists w \in \mathbb{R}^d - \{0\}, w^T A_{k+1} w = 0$.

$$\begin{aligned}
&w^T A_{k+1} w = 0 \\
&\implies w^T(A_{k+1} - cuu^T)w + w^T(cuu^T)w = 0 \\
&\implies w^T(A_{k+1} - cuu^T)w = 0 \land w^T(cuu^T)w = 0 \\
&\implies (\alpha\gamma_k)^T(cuu^T)(\alpha\gamma_k) = 0 \\
&\implies c\alpha^2(\gamma_k^T\delta_k)^2 = 0 && (u = \delta_k) \\
&\implies \alpha^2(\delta_k^T Q \delta_k) = 0 && (\gamma_k = Q\delta_k \text{ and } 2)
\end{aligned}$$

This is not possible because $\delta_k^T Q \delta_k > 0$ (because Q is positive definite) and $\alpha \neq 0$ (because $w \neq 0$). Therefore, we have a contradiction. Therefore, A_{k+1} is positive definite. □

Lemma 6 (Proof omitted (probably beyond scope of course))**.**

$$\forall k \geq 1, \forall i \in [0, k-1], A_k\gamma_i = \delta_i \land \delta_k^T Q \delta_i = 0$$

Let $\Delta = \{\delta_0, \delta_1, \ldots\}$. Lemma 6 states that Δ is Q-conjugate. This implies that Δ is linearly independent. By lemma 3, we get that rank-2 updates converge to minimum in $d + 1$ iterations.

4 BFGS

Instead of modeling the change in hessian's inverse, we'll now model the change in the hessian. But we need to do it in a way such that the change in the inverse is also easy to compute.

Let B_k be an approximation to the hessian and A_k be an approximation to the inverse of the hessian. Then $\gamma_k = B_{k+1}\delta_k$ and $\delta_k = A_{k+1}\gamma_k$.

We'll chose the update rule as

$$B_{k+1} = B_k + cuu^T + bvv^T$$

This will make sure that B_k is symmetric implies B_{k+1} is symmetric.

Applying the Quasi-Newton condition, we get

$$\gamma_k = B_{k+1}\delta_k \implies \gamma_k - B_k\delta_k = (cu^T\delta_k)u + (bv^T\delta_k)v$$

Let $u = \gamma_k$ and $v = B_k\delta_k$.

$$c = \frac{1}{u^T\delta_k} = \frac{1}{\gamma_k^T\delta_k} \qquad\qquad d = \frac{-1}{v^T\delta_k} = \frac{-1}{\delta_k^T B_k\delta_k}$$

$$B_{k+1} = B_k + \frac{\gamma_k^T\gamma_k}{\gamma_k^T\delta_k} - \frac{B_k\delta_k\delta_k^T B_k}{\delta_k^T B_k\delta_k}$$

Similar to theorem 5, we can prove that B_{k+1} is positive definite for quadratic functions. This implies that A_{k+1} is also symmetric and positive definite for quadratic functions.

To invert B_{k+1}, we'll use the Sherman-Morrison formula.

Theorem 7 (Sherman-Morrison formula)**.** *Let A be an invertible matrix. Then $A + uv^T$ is invertible iff $1 + v^TA^{-1}u \neq 0$. Also,*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

Applying the formula twice, we get

$$A_{k+1} = A_k + \frac{\delta_k\delta_k^T}{\delta_k^T\gamma_k}\left(1 + \frac{\gamma_k^TA_k\gamma_k}{\delta_k^T\gamma_k}\right) - \frac{A_k\gamma_k\delta_k^T + \delta_k\gamma_k^TA_k}{\delta_k^T\gamma_k}$$

5 Broyden Family

Let's explore this update rule:

$$A_{k+1} = A_k + a\frac{\delta_k\delta_k^T}{\delta^T\gamma_k} + c\frac{A_k\gamma_k\gamma_k^TA_k}{\gamma_k^TA\gamma_k} - b\frac{A_k\gamma_k\delta_k^T + \delta_k\gamma_k^TA_k}{\delta_k^T\gamma_k}$$

Applying the Quasi-Newton condition, we get

$$\delta_k - A_k\gamma_k = \left(a - b\frac{\gamma_k^TA_k\gamma_k}{\delta_k^T\gamma_k}\right)\delta_k + (c-b)A_k\gamma_k$$

Equating coefficients of δ_k and γ_k, we get

$$a = 1 + b\frac{\gamma_k^TA_k\gamma_k}{\delta_k^T\gamma_k} \qquad\qquad c = b - 1$$

On rearranging, we get

$$A_{k+1} = \left(A_k + \frac{\delta_k\delta_k^T}{\delta_k^T\gamma_k} - \frac{A_k\gamma_k\gamma_k^TA_k}{\gamma_k^TA_k\gamma_k}\right) + b(\gamma_k^TA_k\gamma_k)w_kw_k^T$$

where

$$w = \frac{\delta_k}{\delta_k^T \gamma_k} - \frac{A_k \gamma_k}{\gamma_k^T A_k \gamma_k}$$

This update rule is called the Broyden Family. Note that the first term is the same as the rank-2 update.

Define the following 2 functions:

$$\text{rank-2}(A, \delta, \gamma) = A + \frac{\delta \delta^T}{\delta^T \gamma} - \frac{A \gamma \gamma^T A}{\gamma^T A \gamma}$$

$$\text{bfgs}(A, \delta, \gamma) = A + \frac{\delta \delta^T}{\delta^T \gamma} \left(1 + \frac{\gamma^T A \gamma}{\delta^T \gamma} \right) - \frac{A \gamma \delta^T + \delta \gamma^T A}{\delta^T \gamma}$$

The Broyden family can also be rewritten as

$$A_{k+1} = (1 - b)\, \text{rank-2}(A_k, \delta_k, \gamma_k) + b\, \text{bfgs}(A_k, \delta_k, \gamma_k)$$